OpenEarthMap: A Benchmark Dataset for Global High-Resolution Land Cover Mapping

Junshi Xia[1,*], Naoto Yokoya[2,1,*,†], Bruno Adriano[1,*], and Clifford Broni-Bediako[1]
[1]RIKEN AIP, Japan {junshi.xia,bruno.adriano,clifford.broni-bediako}@riken.jp
[2]The University of Tokyo, Japan yokoya@k.u-tokyo.ac.jp

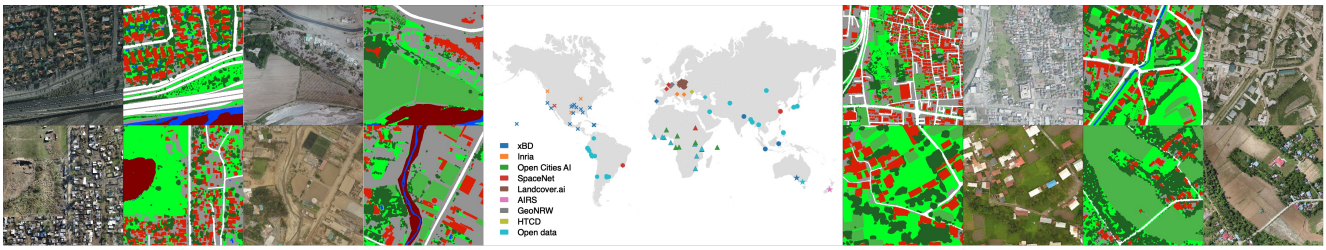


Figure 1: A world map showing the locations of 97 regions included in OpenEarthMap and eight annotated examples.

Abstract

We introduce OpenEarthMap, a benchmark dataset, for global high-resolution land cover mapping. OpenEarthMap consists of 2.2 million segments of 5000 aerial and satellite images covering 97 regions from 44 countries across 6 continents, with manually annotated 8-class land cover labels at a 0.25–0.5m ground sampling distance. Semantic segmentation models trained on the OpenEarthMap generalize worldwide and can be used as off-the-shelf models in a variety of applications. We evaluate the performance of state-of-the-art methods for unsupervised domain adaptation and present challenging problem settings suitable for further technical development. We also investigate lightweight models using automated neural architecture search for limited computational resources and fast mapping. The dataset is available at https://open-earth-map.org.

1. Introduction

Land cover classification maps are the basic information for decision making in various applications, such as land use planning, food security, resource management, and disaster response. Meter-level resolution satellite imagery have been used to map the world, as represented by GlobeLand30 [6], FROM-GLC [5], and recent benchmarks such as OpenSentinelMap [20] and DynamicEarthNet [44]. Satellite imagery at a sub-meter level of ground sampling distance (GSD) enables the extraction of core map information such as buildings and roads. In recent years, there has been substantial progress in automatic construction of building footprints over large areas [39].

Since the advent of deep learning [2], a great deal of effort has been devoted to developing benchmark datasets for high-resolution remote sensing image analysis to facilitate advances in theory and practice. SpaceNet [47] and IEEE GRSS DFC [18], among others, regularly introduce benchmark datasets to the public through competitions that drive research and development. Building detection, road detection, object detection, and land cover classification (semantic segmentation) are the most typical tasks for which these datasets are used in supervised learning [57, 29]. Apart from supervised learning, these datasets have been used in more realistic problems, including transfer learning [49], semi-supervised learning [4] and weakly supervised learning [36, 21]. Benchmark datasets that contribute to solving social problems regarding change detection and disaster damage mapping have been developed as well [14, 16].

Benchmark datasets for semantic segmentation at submeter level resolution have two problems: regional disparity and annotation quality. The regions included in many benchmarks are often biased toward developed countries. Thus, benchmark datasets for regions where map information is not well maintained are scarce. Two main reasons why this problem has not been easily solved are the lack of high-resolution open aerial imagery in developing countries and that commercial high-resolution satellite imagery are basically not redistributable. Other than buildings and roads, the annotation quality of land cover labeling in existing benchmarks is coarse, even though images are at submeter level resolution. This is due to the high cost of manually labeling sub-meter-resolution imagery in spatial detail. Thus, most of the labeling data are based on OpenStreetMap [35] and open map data from local governments.

In this work, we propose OpenEarthMap, a benchmark dataset for global high-resolution land cover mapping with the goal of providing automated mapping for everyone. OpenEarthMap presents a major advance over existing data

*Equal contribution. †Corresponding author.

Table 1: Summary of remote sensing benchmark datasets for semantic segmentation. B: building extraction, R: road extraction, LC: land cover mapping, and CD: change detection. The number of segments was counted on available labels.

Image level	GSD (m)	Dataset	Task	Classes	Countries	Regions	Area (km^2)	Segments
Meter level	10	OpenSentinelMap [20]	LC	15	—	—	505,202	3,467,552
	3	DynamicEarthNet [44]	LC/CD	7	—	75	707	897,855
Sub-meter level	0.3–0.5	SpaceNet 1&2 [47]	B	2	5	5	5,555	685,235
	0.5/0.3/0.5	DeepGlobe [12]	R/B/LC	2/2/7	—	—	2,220/984/1,717	—/302,701/20,697
	0.02–0.2	Open Cities AI [33]	B	2	8	11	419	792,484
	0.5	xBD [16]	B/CD	2/4	15	21	3,382	850,736
	0.3	LoveDA [49]	LC	7	1	3	536	166,768
	0.25–0.5	OpenEarthMap	LC	8	44	97	799	2,205,395

with respect to geographic diversity and annotation quality (see Table 1). OpenEarthMap consists of 8-class land cover labels at a 0.25–0.5m GSD of 5000 images, covering 97 regions from 44 countries across 6 continents. We adopted RGB images of some existing benchmark datasets for building detection and collected additional images for areas not covered by these benchmarks to balance the regional disparities. All images were manually labeled to ensure high-quality annotation. We evaluate the performance of state-of-the-art methods for semantic segmentation and unsupervised domain adaptation tasks and identify problem settings suitable for further technical development. In addition, lightweight models based on automated neural architectural search are investigated for cases where people requiring automated mapping have limited computational resources or for rapid mapping applications such as disaster response.

2. The Dataset

2.1. Source of Imagery

Our strategy is to reuse images from existing benchmark datasets as much as possible and manually annotate new land cover labels. We selected xBD [16], Inria [30], Open Cities AI [33], SpaceNet [47], Landcover.ai [3], AIRS [8], GeoNRW [1], and HTCD [38] datasets based on the condition that the source images are redistributable, the ground sampling distance (GSD) is equal to or less than 0.5m, and the images have geocoordinate information. If there are enough images of a region, which we defined at a scale of province or city, we sampled 50–70 images of that region at a size of 1024×1024 pixels. The number of images from each dataset we adopted was determined based on the diversity and balance of the continents and countries where the images were taken. For countries and regions not covered by the existing datasets, aerial images publicly available in such countries or regions were collected to mitigate the regional gap, which is an issue in most of the existing benchmark datasets. The open data were downloaded from OpenAerialMap [34] and geospatial agencies [15, 32]. See the supplementary for more details of attribution.

In addition to this geographic diversity, our dataset includes a mixture of images taken from different platforms, including satellite, aircraft, and UAV. For very high-resolution images with GSD less than 0.25m, we resampled the images to 0.3m or 0.5m to account for object size and visual interpretability of the captured area. Basically, the images were selected by a combination of random sampling and manual checking for each region. Moreover, if the number of images of a particular region is very large in the source benchmark dataset, we trained a segmentation model using sequentially labeled data (e.g., every 10 images) and another regression model to estimate the loss. Then, we added the images that have high values of predicted loss, as they are more difficult by a model trained with the available labels to segment.

In the end, we collected a total of 5000 images from 97 regions of six continents. Figure 1 shows annotated samples and the geographic distribution of the 97 regions with different colors indicating the source datasets. Figure 2 depicts the number of images in our dataset for each of the six continents, colored to indicate the origin of the images. Asia, Africa, and South America are not well covered by the source datasets; thus, we added many images from public data to balance the regional disparities. Figure 3 presents a t-SNE 2D plot based on the similarity of image features for the 97 regions. For each region, we used the average of features extracted by EfficientNet-B4 trained as an encoder of U-Net on OpenEarthMap. The 12 representative images in the 2D plot show that different locations correspond to diverse images. It can also be seen that the different source datasets are complementary to each other, and that the diversity of images is enriched by the open data we added. The different symbols correspond to the six continents and enable the similarities between the continents to be seen. For example, regions in Europe and North America as well as Africa and South America are similar in the image features.

2.2. Classes, Annotations, and Data Split

Classes: We provide annotations with eight classes: *bareland*, *rangeland*, *developed space*, *road*, *tree*, *water*, *agriculture land*, and *building*. The class selection is consistent with existing products and benchmark datasets (e.g., LoveDA [49] and DeepGlobe [12]) with sub-meter GSD.

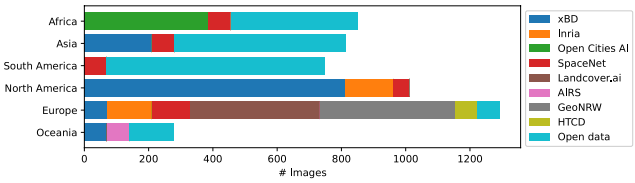


Figure 2: The number of images of the six continents in OpenEarthMap.

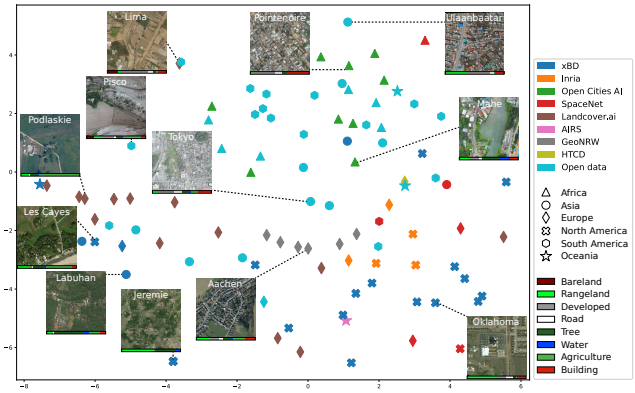


Figure 3: t-SNE 2D visualization of the 97 regions based on features extracted with EfficientNet-B4 trained on OpenEarthMap. The images are samples of 12 regions with horizontal bar charts of class proportions attached at the bottom.

Table 2 shows the number and proportion of labeled pixels and the number of segments of each class. Here, as well as in Table 1, we refer to a segment as a set of connected pixels with the same label, and it was counted using OpenCV's findContours function. It can be seen that the elevated objects (e.g., *tree* and *building*) are finely annotated compared to the ground objects (e.g., *agriculture land*). As can be seen in the horizontal bar charts of 12 representative regions in Figure 3, the class proportions in the different regions are diverse.

Annotations: A total of 16 people worked on the annotation process: 8 people were responsible for annotating the images, while the remaining 8 people performed quality checks to point out errors. One person labels an image and at least two people perform the quality check. We spent a longer time labeling the first 100 images and exchanging ideas with each other to ensure that all participants were in agreement about the class definitions. On average, the labeling took 2.5 hours per image. This is significantly longer than the 1.5 hours of Cityscapes [11], which illustrates the difficulty of labelling remote sensing images. All the labeling was done manually. For the labeling of images of the existing benchmark datasets, only the *building* class was used as the starting point. However, since a lot of label noise was found, segments of *buildings* were also manually modified. The most important feature of OpenEarthMap's labeling is its level of spatial detail. As shown in Table 1, the area covered by the images in OpenEarthMap is not very large compared to the other benchmark datasets, however, the number of segments is 10 times more than that of LoveDA.

Table 2: The number and proportion of pixels and the number of segments of the eight classes.

Color (HEX)	Class	Pixels		Segments
		Count (M)	(%)	(K)
800000	Bareland	74	1.5	6.3
00FF24	Rangeland	1130	22.9	459.4
949494	Developed space	798	16.1	382.7
FFFFFF	Road	331	6.7	27.9
226126	Tree	996	20.2	902.9
0045FF	Water	161	3.3	18.7
4BB549	Agriculture land	680	13.7	18.2
DE1F07	Building	770	15.6	389.3

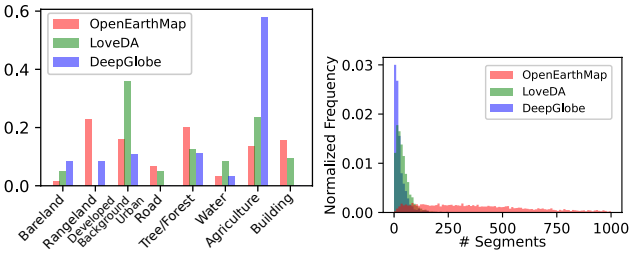
The accuracy of human annotations is evaluated by having two different people labeling 200 images twice. We selected two or three images with as many classes as possible based on the first annotation from each region to constitute the 200 images. The percentage of pixels that were labeled as the same class in the two different annotations by different people is 78%. This percentage is significantly lower than the 96% in Cityscapes [11], suggesting that annotation of high-resolution remote sensing images is much more challenging than annotation of urban street scenes. The relationship between human labeling accuracy and estimation accuracy of the state-of-the-art segmentation models is discussed in Section 3.4.

Data split: For the semantic segmentation task, the images from each region were randomly divided into training, validation, and test sets with a ratio of 6:1:3, which respectively yielded 3000, 500, and 1500 images out of the total 5000 images. To ensure that all classes in each region are included in the training set and as many classes as possible are included in the test set, the split with the least mismatch between the training and test classes was selected from multiple random trials. For the unsupervised domain adaptation (UDA) tasks, we adopt two ways of data split to investigate regional-level and continent-wise domain gaps. For regional-level UDA, the entire dataset is divided into 73 and 24 regions for source and target domains, respectively. The split was performed in such a way that both the source and the target domains consist of relatively even distribution of the countries from all six continents as well as a balance between urban and rural areas. This split is not as extreme as the urban-rural split in LoveDA but rather it is a realistic scenario in domain adaptation where OpenEarthMap is at hand as source data and adapts models for mapping in any new region, not only urban-rural adaptation. For continent-wise UDA, we use data from one continent as the source domain and other continents as the target domains.

2.3. Comparison with Related Datasets

Very recently, meter-level resolution benchmarks have made great progress in global land cover mapping;

OpenSentinelMap [20] is featured in its comprehensive coverage of the globe exploiting open data of Sentinel-2 and OpenStreetMap while DynamicEarthNet [44] is advantageous at high-temporal resolution. OpenEarthMap goes one step further in providing spatially detailed annotation at the sub-meter level. A more detailed comparison is made with LoveDA [49] and DeepGlobe [12], which have similar resolution and class definitions as OpenEarthMap. Figure 4a shows a comparison of the class proportions of the three datasets. It should be noted that LoveDA does not include *rangeland*, and that in the DeepGlobe dataset for land cover classification, *buildings* and *roads* are included in the *urban* class. There is no dominant class in OpenEarthMap and the class proportions are relatively balanced. The normalized histogram of the number of segments in a single image is shown in Figure 4b. In terms of image size, LoveDA is the same (1024×1024 pixels) as OpenEarthMap, while DeepGlobe is larger (2448×2448 pixels). The histogram of OpenEarthMap has a very long tail, showing a much larger number of segments in each image of OpenEarthMap than the other datasets. The spatially detailed labeling of the OpeneEarthMap is reflected in the cross-dataset evaluation and the out-of-sample prediction results of trained models presented in Sections 5 and 6.



(a) Class proportions (b) Histogram of # segments

Figure 4: (a) The proportions of annotated pixels per class and (b) Normalized histograms of the number of segments for OpenEarthMap, LoveDA and DeepGlobe datasets.

3. Land Cover Semantic Segmentation

3.1. Baselines

For the land cover semantic segmentation task, CNN-based and Transformer-based architectures were evaluated and compared on the OpenEarthMap dataset. More specifically, the chosen models are U-Net [37], U-NetFormer [50], FT-U-NetFormer [50], DeepLabV3 [7], HRNet [41], SETR [56], SegFormer [54], and UPerNet [53] with backbones of ViT [13], Twins [10], Swin Transformer [25], ConvNeXt [26], and K-Net [55].

3.2. Results

General results: The results obtained on the test set of OpenEarthMap are presented in Table 3. The main findings are discussed as follows: (1) U-Net with EfficientNet-B4 as backbone outperforms both U-Net with ResNet-34 and U-Net with VGG-11. The reason might be that EfficientNet-B4 is more effective for extracting relevant features, and to that effect, both high-level features and low-level spatial information are used for robust segmentation. (2) UPerNet with Swin-B and Twins, as well as SegFormer and K-Net perform better than DeepLabV3 and HRNet. This might be attributed to the strong modeling capabilities and dynamic feature aggregation of Swin-B, Twins, and MiT-B5. (3) U-NetFormer and FT-U-NetFormer share the top positions because both methods adopt a global–local Transformer block to construct global and local information in the decoder, and use advanced encoder (e.g., ResNeXt and Swin-B) to extract features. (4) UPerNet with ViT and ConvNeXt, along with SETR obtain worse results than other Transformer-based models. Two reasons might be that the hyperparameters (e.g., optimizer and learning rate) of these methods may need to be carefully tuned, and advanced data augmentation may be required for transfer learning from ImageNet to the OpenEarthMap dataset. In all, considering performance along with the number of parameters and FLOPs, U-Net-EfficientNet-B4, UPerNet-Swin-B, and FT-U-NetFormer are recommended.

Visualization: Examples of segmentation results obtained from some selected methods are presented in Figure 5. The U-Net-EfficientNet-B4 and FT-U-NetFormer produce the best detailed visualization results. In the first row of Figure 5, DeeplabeV3 wrongly classified the *water* area of the dam as *rangeland* while other methods identified them. In the second row, U-Net-EfficientNet-B4, SegFormer and FT-U-NetFormer were able to identify the tiny roads in the top-right parts of the image. *Water* and *bareland* classes respectively achieved the highest and the lowest accuracies in all methods. The boundaries of the *buildings* and the *roads* were difficult to identify properly because of disorganized layouts and varying sizes. *Rangeland*, *agricultural land* and *trees* are easy to confuse due to the similarities in their spectra. *Roads* were easily misclassified as *developed space* because parking lots and cover materials in some rural areas are quite similar.

3.3. Neural Architecture Search

LoveDA [49], DeepGlobe [12], and other previous benchmarks [3, 30, 8] were experimented with only manually designed networks [37, 41, 7, 27, 23] for the semantic segmentation task. In contrast, we further experimented the OpenEarthMap dataset with two automated neural architecture search methods, SparseMask [52] and FasterSeg [9], by automatically searching for compact segmentation architectures. Such architectures might offer a useful baseline for research in the field of automated neural architecture search in remote sensing with OpenEarthMap. Following the ar-

Table 3: Semantic segmentation results of the baseline models on the test set of the OpenEarthMap dataset. The results are based on test-time augmentation (TTA), in particular flipping.

Method	Backbone	IoU (%)								mIoU	Params	FLOPs
		Bareland	Rangeland	Developed	Road	Tree	Water	Agriculture	Building	(%)	(M)	(G)
U-Net	VGG-11	40.69	56.76	53.99	62.16	72.44	82.81	73.14	77.77	64.97	**18.26**	233.33
U-Net	ResNet-34	40.35	57.75	54.92	62.87	72.65	82.24	74.06	78.58	65.43	24.44	126.68
U-Net	EfficientNet-B4	50.63	58.17	56.27	64.83	73.20	86.02	76.28	80.20	68.20	20.30	**45.47**
U-NetFormer	ResNeXt101	46.09	60.67	**58.12**	65.07	**73.77**	86.34	76.98	79.96	68.37	192.71	769.25
FT-U-NetFormer	Swin-B	**50.19**	**60.84**	57.58	**65.85**	73.33	**87.44**	**77.50**	**80.29**	**69.13**	95.98	498.37
DeepLabV3	ResNet-50	39.11	56.16	52.28	60.57	71.25	79.32	70.75	75.83	63.16	68.14	269.76
HRNet	W48	39.71	55.50	53.49	59.22	71.10	79.03	71.38	75.12	63.07	65.89	94.06
UPerNet	ViT	34.39	54.45	50.64	54.57	69.73	79.24	66.22	74.92	60.52	144.17	395.07
UPerNet	Swin-B	44.52	58.98	54.78	63.43	72.20	83.71	72.97	78.11	66.09	59.94	236.08
SegFormer	MiT-B5	36.84	57.94	53.53	63.60	70.51	80.11	72.21	77.35	64.01	81.97	51.86
SETR PUP	ViT-L	45.35	55.72	51.31	55.47	67.63	73.12	67.14	75.48	61.40	309.35	212.45
UPerNet	Twins	37.29	57.62	53.83	60.23	72.32	81.93	71.71	77.49	64.05	90.96	250.91
UPerNet	ConvNeXt	40.61	54.94	51.76	58.47	70.44	75.95	68.94	74.30	61.93	122.1	292.42
K-Net	Swin-B	44.02	57.81	54.85	62.91	71.76	85.18	73.41	78.91	66.11	246.97	419.51

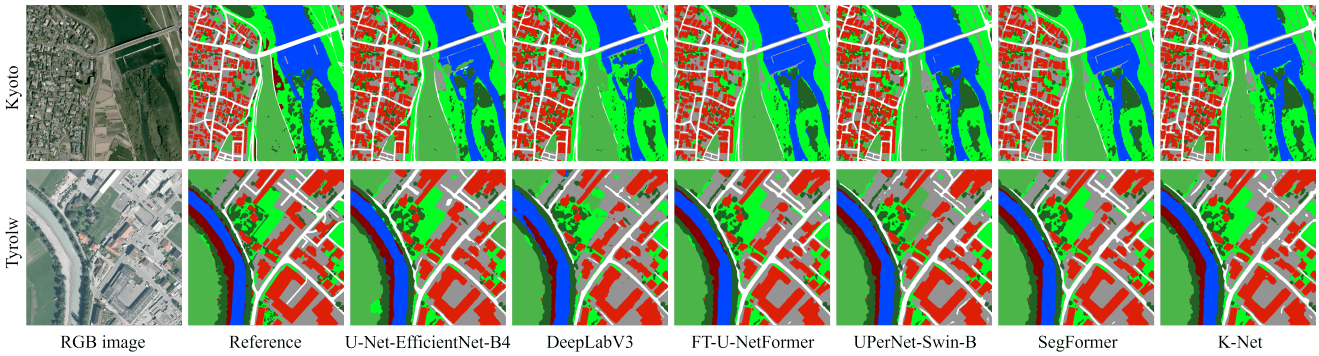


Figure 5: Visual comparison of land cover mapping results of some of the baseline models presented in Table 3.

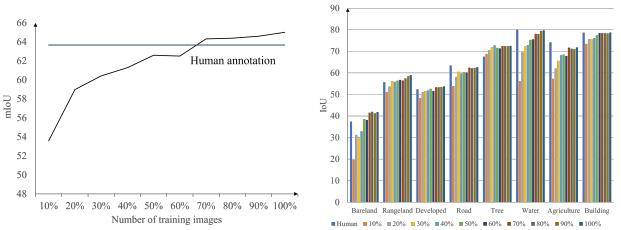
Table 4: Lightweight models discovered on OpenEarthMap training set. FLOPs and FPS are measured on 1024×1024 input, and mIoU on the test set of OpenEarthMap.

Method	Trial	Params (M)	FLOPs (G)	FPS (ms)	mIoU (%) No TTA	mIoU (%) TTA
SparseMask	1st	2.96	**10.28**	51.2	58.23	**60.21**
	2nd	3.10	10.39	52.2	58.06	60.00
FasterSeg	1st	**2.23**	14.58	143.2	57.55	58.35
	2nd	3.47	15.37	**171.3**	**58.51**	59.41

chitecture search protocols in both methods (see the supplementary for more details), we searched for lightweight segmentation networks on the OpenEarthMap dataset. Four experiments were performed, two with each method, and the results are presented in Table 4. Both methods were able to discover compact networks, however, FasterSeg discovered the lightest-weight network. The networks discovered by SparseMask have less computational complexity but with low inference speed. Whereas FasterSeg networks have high computation cost and high inference speed. For real-time mapping (no TTA), FasterSeg might serve as a baseline for the OpenEarthMap dataset. For non real-time mapping (where TTA is used), SparseMask might be adopted as a baseline. Compared to the manually designed baseline models presented in Table 3, the lightweight discovered networks ($< 4M$ params) competed with UPerNet-ViT ($144.17M$ params) and trailed behind FT-U-NetFormer ($95.98M$ params) by approximately 9% accuracy rate.

3.4. Human Annotation *vs* Machine Prediction

As mentioned in Section 2.2, 200 images were labeled twice by different people. The remaining 4800 images were used to train UPerNet with Swin-B to compare the quality of human labeling with the results from the machine. To effectively investigate the comparison, the number of training images was varied from 10% to 100%; the results are shown in Figure 6. It can be seen that with 50% of the training images, the machine attains almost the same level of human annotation and larger training percentages improve the accuracy (see Figure 6a). For human annotation, the challenging classes include *bareland*, *rangeland*, and *tree*. For *bareland*, *rangeland*, *developed space*, and *tree* classes, 50%, 30%, 50%, and 10% of the training set, respectively, yielded better results than the ones of human annotation (see Figure 6b). The challenging class for the machine is *agriculture land*, where it trails behind human annotation by 2.3%. Regarding *road*, *water*, and *building* classes, with 100% of the training images, the machine slightly ($< 0.34\%$) trails behind the human annotation.



(a) mIoU (b) Class-specific IoUs

Figure 6: Human annotation *vs* machine predictions with varying numbers of images from the training set. Human annotation accuracies mean the IoUs between two different human annotations.

3.5. Learning from Limited Labels

We also investigated the performance of CNN-based (U-Net-EfficientNet-B4) and Transformer-based (SegFormer, UPerNet-Swin-B and K-Net) models on limited training samples. Table 5 presents the results of using only 10% of the OpenEarthMap training set to train the models. It is apparent from Table 5 that U-Net-EfficientNet-B4 outperforms all three Vision Transformer-based methods in all the class-specific IoUs by about 6-15%. The main reason is that the representation capacity of ViTs typically lacks the inductive bias in CNNs. Therefore, ViTs require more training data than CNNs [24, 13]. We believe that Vision Transformers with small-sized data [24] or limited labels [19] is an interesting topic that requires further study. Moreover, data augmentation, regularization, and tuning of hyper-parameters still need to be explored when training on limited training data [40].

4. Unsupervised Domain Adaptation

4.1. Baselines

For the unsupervised domain adaptation task, a metric-based method (MCD [46]), adversarial training methods including AdaptSeg [45], category-level adversarial network (CLAN) [28], TransNorm [51], and fine-grained adversarial learning framework for domain adaptive (FADA) [48]), as well as self-training methods including pyramid curriculum DA (PyCDA) [22], class-balanced self-training (CBST) [58], instance adaptive self-training (IAST) [31], and DAFormer [17] are adopted. DAFormer is based on SegFormer and the others are based on DeepLabV2.

4.2. Results

Regional-level UDA: We investigated the regional-level domain gap since different regions in the same continent might suffer from a distribution shift. The results obtained on the test set of 24 regions of OpenEarthMap are presented in Table 6. In general, the Oracle settings obtained the best results. Due to the regional domain gap, the source-only settings yielded the lowest accuracy. The results of

Table 5: Semantic segmentation results of selected baseline models trained on only 10% of OpenEarthMap training set.

	Bare	Range	Dev	Road	Tree	Water	Agri	Building	mIoU
U-Net-EfficientNet-B4	**32.62**	**52.43**	**49.77**	**58.47**	**69.26**	**74.39**	**70.16**	**74.35**	**60.18**
SegFormer	16.15	44.08	45.88	51.39	65.72	61.42	58.54	69.71	51.61
UPerNet-Swin-B	18.32	47.82	48.2	53.46	66.89	59.62	55.22	69.55	52.39
K-Net	18.62	50.26	48.93	55.22	66.45	60.76	62.06	72.33	54.33

Table 6: Unsupervised domain adaptation results obtained on the test set of 24 regions in the OpenEarthMap dataset.

	Type	IoU (%)								mIoU (%)
		Bare	Range	Dev	Road	Tree	Water	Agri	Build	
DeeplabV2-based										
Oracle	—	37.06	43.65	38.03	43.12	61.61	73.89	75.90	63.93	54.65
Source only	—	26.86	42.14	36.48	42.03	58.58	61.35	70.77	61.87	50.01
MCD	—	16.77	41.55	35.89	44.24	56.15	57.84	62.57	63.83	47.36
AdaptSeg	AT	28.77	41.47	36.09	45.16	46.65	34.48	68.47	63.74	45.60
FADA	AT	26.29	37.91	34.91	37.13	54.19	40.68	65.36	58.32	44.35
CLAN	AT	22.90	42.25	39.49	44.12	58.98	58.99	59.51	64.53	48.85
TransNorm	AT	27.54	45.13	37.99	45.56	57.06	63.84	66.26	64.71	51.01
PyCDA	ST	21.95	32.33	22.89	34.81	44.95	34.16	56.74	55.31	37.89
CBST	ST	29.64	43.79	37.99	49.19	57.33	60.75	71.93	65.46	52.01
IAST	ST	33.68	43.64	37.03	45.16	59.61	72.08	74.72	61.77	53.46
SegFormer-based										
Oracle	—	43.14	53.02	51.50	61.13	68.06	81.89	81.38	79.81	64.99
Source only	—	28.37	48.96	46.49	54.05	67.62	75.32	77.93	75.79	59.32
DAFormer	ST	**37.16**	**51.07**	**50.36**	**58.07**	**68.34**	**78.39**	**78.08**	**77.30**	**62.35**

source-only SegFormer are significantly better than those of source-only DeepLabV2. Compared to manufactured classes (i.e., *building* and *road*), the accuracies of natural classes (i.e., *water* and *bareland*) decreased significantly. With the exception of TransNorm, the adversarial training methods did not perform well on this task due to the diversity in the OpenEarthMap dataset. TransNorm slightly improved the performance because the source and the target images have distinct spectral statistics since they were taken from different sensors and regions. The class imbalance problem is addressed using pseudo-label creation via the CBST and IAST techniques, resulting in higher performance. Due to better domain generalization of SegFormer and effective training strategy in self-training, DAFormer obtained the best mIoU of 62.35%. Visual examples of the UDA results are presented in Figure 7. In the first row of Figure 7, source-only DeepLabV2 can barely identify the *water* area (top-right) and the roads (bottom-right). IAST and CBST performance improves for *water* but they lose the ability to recognize the roads. DAFormer performs very well in the two complex areas. In the second row, DAFormer shows better visualization results in the small *water* area (top-right) and the boundaries of *roads* and *buildings* than the other UDA methods.

Continent-wise UDA: We also investigated the continent-wise domain gap on the OpenEarthMap dataset using U-Net-EfficientNet-B4, SegFormer, and DAFormer. The results are presented in Figure 8. Compared to the UDA settings (e.g., GTA5→Cityscapes with similar content and different style) in computer vision and previous settings in remote sensing (e.g., urban→rural in LoveDA), UDA on continent-wise has larger content and style gaps. The lim-

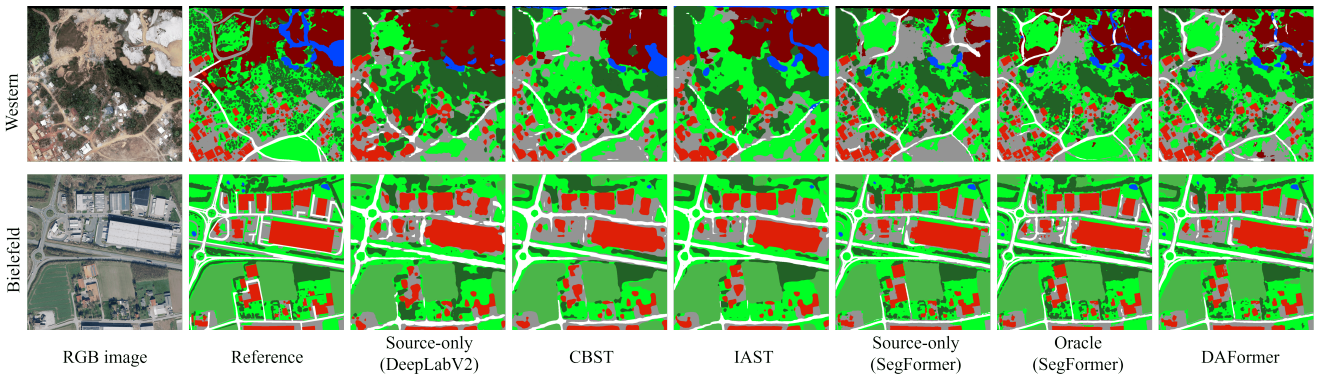


Figure 7: Visual comparison of unsupervised domain adaption results of some of the baseline models presented in Table 6.

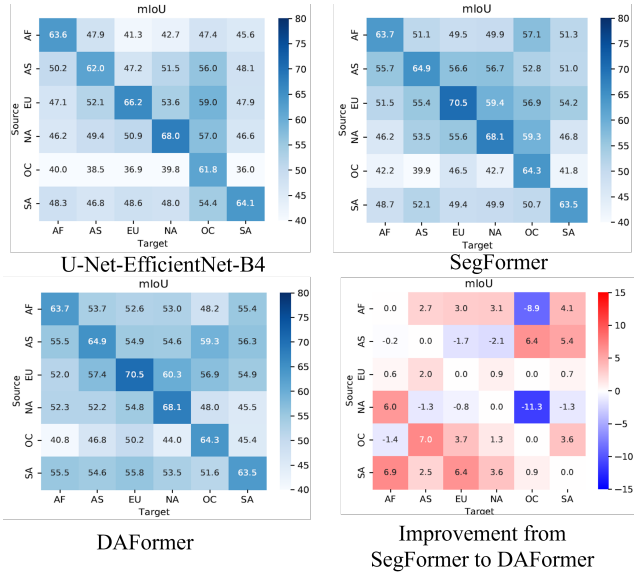
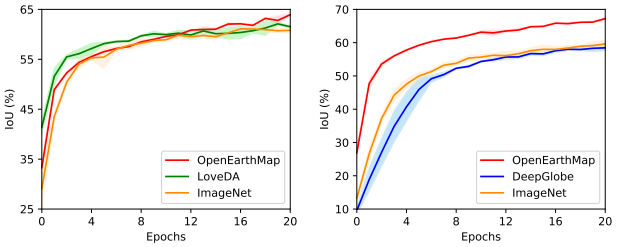


Figure 8: Continent-wise UDA results. Asia: AS, Europe: EU, Africa: AF, North America: NA, South America: SA, and Oceania: OC.

ited data of Oceania (OC) led to the lowest transferred results when OC is treated as the source domain. In contrast, the performance with OC as the target domain is better than other settings. Expect OC, U-Net-EfficientNet-B4 and SegFormer indicated two minor domain gaps: Europe (EU)-to-North America (NA) and Asia (AS)-to-NA. The most prominent domain gap revealed by EfficientNet-B4 and SegFormer is Africa (AF)-to-EU and NA-to-AF, respectively. For challenging UDA settings, SegFormer is generally better than U-Net-EfficientNet-B4 (26 out of 30), which is the opposite of the results in the semantic segmentation (see Table 3) and the regional UDA setting (see Table 6). DAFormer improved the results compared to SegFormer in many cases (20 out of 30). DAFormer on AF-to-OC and NA-to-OC achieved significantly poor results due to the collapsed construction of pseudo labels in the limited data of OC. Thus, challenging continent-wise UDA settings are worth exploring and possible solutions may include the



(a) Fine-tuning on DeepGlobe (b) Fine-tuning on LoveDA

Figure 9: Comparison among OpenEarthMap, LoveDA and DeepGlobe pre-trained models.

extension of DAFormer or new UDA method with U-Net-EfficientNet-B4.

5. Cross-Dataset Evaluation

In this section, we evaluate the advantage of using the OpenEarthMap dataset as a starting point (fine-tuning) in the semantic segmentation task over other open-source land cover mapping datasets. Here we compare OpenEarthMap with LoveDA [49] and DeepGlobe [12]. We adopted the same U-Net model with an EfficientNet-B4 as a backbone listed in Table 3 and trained from scratch on the three datasets using similar training settings as Section 3.2. Then, we fine-tuned the OpenEarthMap and the LoveDA pre-trained models on the DeepGlobe dataset. Similarly, the OpenEarthMap and the DeepGlobe pre-trained models were fine-tuned on the LoveDA dataset. All the experiments were run threefold; we report the mean and the standard deviation segmentation accuracy for 20 epochs. As presented in Figure 9, the results indicate that using a model that is pre-trained on OpenEarthMap as a starting point could yield better performance than models pre-trained on LoveDA and DeepGlobe. For example, when fine-tuned on the DeepGlobe dataset, the initial IoU score of the OpenEarthMap pre-trained model is about 4% higher than the fully-trained model on DeepGloble (see Figure 9a). Although the OpenEarthMap pre-trained model is slightly lower than the LoveDA pre-trained one in early epochs, OpenEarth-

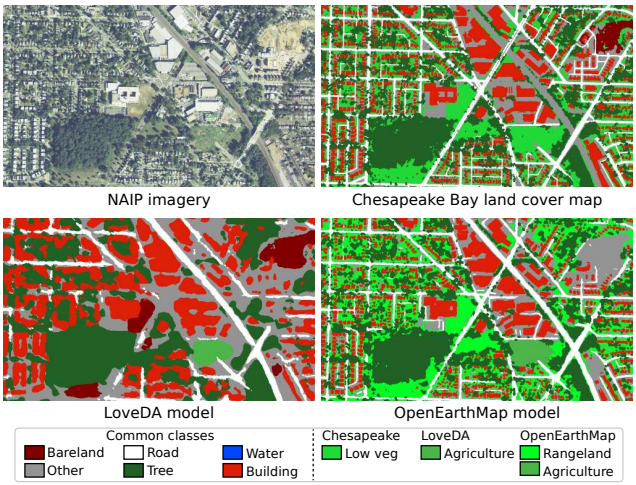


Figure 10: Visual comparison of Chesapeake Bay land cover map with land cover maps generated by U-Net models trained on LoveDA and OpenEarthMap. The NAIP image is the source data.

Map increasingly outperforms both models as the number of epochs increases. Furthermore, when fine-tuned on the LoveDA dataset, the OpenEarthMap pre-trained model attains a more than 20% increase in the initial IoU score, and its performance remains higher when the number of epochs increases (see Figure 9b).

6. Demonstration on Out-of-Sample Imagery

To further investigate the generalization performance of a model trained on OpenEarthMap, we created land cover classification maps from out-of-sample imagery (i.e., images that are not included in OpenEarthMap). See the supplementary for more results. Here we present a map created from an NAIP [43] image resampled at 0.5m GSD. A Chesapeake Bay land cover map [42] was used as reference to evaluate performance. The Chesapeake Bay land cover map consists of 13 classes. To fairly compare the mapping results of the Chesapeake Bay land cover tool with those produced by U-Net models trained on OpenEarthMap and LoveDA, we adopted six common classes (*bareland*, *other*, *road*, *tree*, *water*, and *building*) among the datasets and performed quantitative evaluation. Table 7 shows the accuracy of the land cover mapping for an area of approximately $15km \times 28km$ in US, spanning from Washington, DC to Maryland. The IoUs from OpenEarthMap model are significantly higher than those from LoveDA, and the scores are high enough for practical mapping except *bareland*. The accuracy of *bareland* is low due to inconsistency in class definitions. For example, in the Chesapeake Bay land cover map, a construction site is labeled as *bareland*, while OpenEarthMap labels the same area as *developed space*. Figure 10 shows a visual example of the mapping results. Note that unlike the quantitative evaluation in Table 7, the vegetation classes (*low vegetation*, *agriculture land*, and *rangeland*) that differ among the datasets are visualized in different colors. The OpenEarthMap model result is similar to the Chesapeake Bay land cover map in both classification and resolution, and achieved very fine spatial segmentation compared to the LoveDA model. This demonstrates the advantage of OpenEarthMap over LoveDA and how finely OpenEarthMap's annotations are spatially detailed.

Table 7: Generalization performance of models trained on OpenEarthMap and LoveDA, and evaluated with IoU (%) using the Chesapeake Bay high-resolution land cover map.

Dataset	Bare	Other	Road	Tree	Water	Build	mIoU
OpenEarthMap	**9.29**	**58.27**	**49.29**	**75.72**	**85.46**	**63.44**	**56.91**
LoveDA	3.07	40.14	37.71	69.34	80.12	45.85	46.04

7. Conclusion and Societal Impacts

The existing benchmarks for land cover classification at sub-meter resolution lack regional diversity and annotation quality. To address this problem, we introduce OpenEarthMap, a benchmark dataset, for global high-resolution land cover mapping. The diversity of the dataset is shown in the coverage of 97 regions from 44 countries across 6 continents, while its finely detailed annotations are reflected in the generalization of the feature space. To demonstrate the practical usefulness of OpenEarthMap, we perform baseline experiments with several state-of-the-art models for semantic segmentation and UDA tasks, and create land cover maps for out-of-sample imagery to show that models trained on OpenEarthMap can adapt and generalize across the globe. We also demonstrate the challenges of the continent-wise domain gap and limited data training. We experiment NAS-based lightweight models for mapping with resource-limited devices. Further technical development is needed to improve the performance in continent-wise domain adaptation, limited training data, and lightweight models on OpenEarthMap for worldwide evaluation. The dataset is made publicly available for other researchers to build on it and create new practical tasks.

Societal Impacts: OpenEarthMap models could enable automated mapping of any location on Earth, which can support decision making in disaster response, environmental conservation, and urban planning. However, such models will make it easy for anyone to access map information related to national security as well as privacy if sub-meter resolution images are available. Appropriate data analysis ethics and data policies are required to avoid security and privacy breaches.

Acknowledgement

This work was supported by JST FOREST Grant Number JPMJFR206S, Japan.

References

[1] Gerald Baier, Antonin Deschemps, Michael Schmitt, and Naoto Yokoya. Synthesizing optical and sar imagery from land cover maps and auxiliary raster data. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2021.

[2] Yoshua Bengio, Yann Lecun, and Geoffrey Hinton. Deep learning for ai. *Commun. ACM*, 64(7):58–65, 2021.

[3] Adrian Boguszewski, Dominik Batorski, Natalia Ziemba-Jankowska, Tomasz Dziedzic, and Anna Zambrzycka. Landcover. ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1102–1110, 2021.

[4] Javiera Castillo-Navarro, Bertrand Le Saux, Alexandre Boulch, Nicolas Audebert, and Sébastien Lefèvre. Semi-supervised semantic segmentation in earth observation: The minifrance suite, dataset analysis and multi-task network study. *Machine Learning*, pages 1–36, 2021.

[5] B Chen, B Xu, Z Zhu, C Yuan, H Ping Suen, J Guo, N Xu, W Li, Y Zhao, JJSB Yang, et al. Stable classification with limited sample: Transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017. *Sci. Bull*, 64:370–373, 2019.

[6] Jun Chen, Jin Chen, Anping Liao, Xin Cao, Lijun Chen, Xuehong Chen, Chaoying He, Gang Han, Shu Peng, Miao Lu, et al. Global land cover mapping at 30 m resolution: A pok-based operational approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, 103:7–27, 2015.

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[8] Qi Chen, Lei Wang, Yifan Wu, Guangming Wu, Zhiling Guo, and Steven L Waslander. Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings. *ISPRS journal of photogrammetry and remote sensing*, 147:42–55, 2019.

[9] Wuyang Chen, Xinyu Gong, Xianming Liu, Qian Zhang, Yuan Li, and Zhangyang Wang. Fasterseg: Searching for faster real-time semantic segmentation. In *International Conference on Learning Representations*, 2020.

[10] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting spatial attention design in vision transformers. In *NeurIPS*, 2021.

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[12] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 172–181, 2018.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[14] Aito Fujita, Ken Sakurada, Tomoyuki Imaizumi, Riho Ito, Shuhei Hikosaka, and Ryosuke Nakamura. Damage detection from aerial images via convolutional neural networks. In *2017 Fifteenth IAPR international conference on machine vision applications (MVA)*, pages 5–8, 2017.

[15] Geospatial Information Authority of Japan. https://maps.gsi.go.jp/development/ichiran.html. Accessed on 2022-06-28.

[16] Ritwik Gupta, Bryce Goodman, Nirav Patel, Ricky Hosfelt, Sandra Sajeev, Eric Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew Gaston. Creating xbd: A dataset for assessing building damage from satellite imagery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 10–17, 2019.

[17] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *CVPR*, 2022.

[18] IEEE GRSS: Image Analysis and Data Fusion. https://www.grss-ieee.org/technical-committees/image-analysis-and-data-fusion/. Accessed on 2022-06-28.

[19] Zi-Hang Jiang, Qibin Hou, Li Yuan, Daquan Zhou, Yujun Shi, Xiaojie Jin, Anran Wang, and Jiashi Feng. All tokens matter: Token labeling for training better vision transformers. In *NeurIPS*, 2021.

[20] Noah Johnson, Wayne Treible, and Daniel Crispell. Opensentinelmap: A large-scale land use dataset using openstreetmap and sentinel-2 imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1333–1341, June 2022.

[21] Zhuohong Li, Fangxiao Lu, Hongyan Zhang, Lilin Tu, Jiayi Li, Xin Huang, Caleb Robinson, Nikolay Malkin, Nebojsa Jojic, Pedram Ghamisi, et al. The outcome of the 2021 ieee grss data fusion contest—track msd: Multitemporal semantic change detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:1643–1655, 2022.

[22] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In *ICCV*, 2019.

[23] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017.

[24] Yahui Liu, Enver Sangineto, Wei Bi, Nicu Sebe, Bruno Lepri, and Marco De Nadai. Efficient training of visual transformers with small datasets. In *NeurIPS*, 2021.

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer:

Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

[26] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022.

[27] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[28] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *CVPR*, 2019.

[29] Lei Ma, Yu Liu, Xueliang Zhang, Yuanxin Ye, Gaofei Yin, and Brian Alan Johnson. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS journal of photogrammetry and remote sensing*, 152:166–177, 2019.

[30] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3226–3229. IEEE, 2017.

[31] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *ECCV*, 2020.

[32] National Center for Disaster Risk Estimation, Prevention and Reduction (CENEPRED) of Peru. https://sigrid.cenepred.gob.pe. Accessed on 2022-06-28.

[33] Open Cities AI Competition: Segmenting Buildings for Disaster Resilience. https://www.drivendata.org/competitions/60/building-segmentation-disaster-resilience/. Accessed on 2022-06-29.

[34] OpenAerialMap. https://openaerialmap.org/. Accessed on 2022-06-28.

[35] OpenStreetMap. https://www.openstreetmap.org. Accessed on 2022-06-28.

[36] Caleb Robinson, Kolya Malkin, Nebojsa Jojic, Huijun Chen, Rongjun Qin, Changlin Xiao, Michael Schmitt, Pedram Ghamisi, Ronny Hänsch, and Naoto Yokoya. Global land-cover mapping with weak supervision: Outcome of the 2020 ieee grss data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:3185–3199, 2021.

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[38] Ruizhe Shao, Chun Du, Hao Chen, and Jun Li. Sunet: Change detection for heterogeneous remote sensing images from satellite and uav using a dual-channel fully convolution network. *Remote Sensing*, 13(18):3750, 2021.

[39] Wojciech Sirko, Sergii Kashubin, Marvin Ritter, Abigail Annkah, Yasser Salah Eddine Bouchareb, Yann Dauphin, Daniel Keysers, Maxim Neumann, Moustapha Cisse, and John Quinn. Continental-scale building detection from high resolution satellite imagery. *arXiv preprint arXiv:2107.12283*, 2021.

[40] Andreas Peter Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022.

[41] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

[42] The Chesapeake Bay Land Cover Map. https://chesapeake.usgs.gov/phase6/map/#map=7/-8582732.74/4851421.17/0.0/0,4. Accessed on 2022-06-29.

[43] The National Agriculture Imagery Program (NAIP). https://naip-usdaonline.hub.arcgis.com/. Accessed on 2022-06-29.

[44] Aysim Toker, Lukas Kondmann, Mark Weber, Marvin Eisenberger, Andrés Camero, Jingliang Hu, Ariadna Pregel Hoderlein, Çağlar Şenaras, Timothy Davis, Daniel Cremers, Giovanni Marchisio, Xiao Xiang Zhu, and Laura Leal-Taixé. Dynamicearthnet: Daily multi-spectral satellite dataset for semantic change segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21158–21167, June 2022.

[45] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 2018.

[46] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *CoRR*, abs/1412.3474, 2014.

[47] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018.

[48] Haoran Wang, Tong Shen, Wei Zhang, Lingyu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In *ECCV*, 2020.

[49] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021.

[50] Libo Wang, Rui Li, Ce Zhang, Shenghui Fang, Chenxi Duan, Xiaoliang Meng, and Peter M. Atkinson. Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:196–214, 2022.

[51] Ximei Wang, Ying Jin, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Transferable normalization: Towards improving transferability of deep neural networks. In *NeurIPS*, 2019.

[52] Huikai Wu, Junge Zhang, and Kaiqi Huang. Sparsemask: Differentiable connectivity learning for dense image prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[53] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018.

[54] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021.

[55] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-Net: Towards unified image segmentation. In *NeurIPS*, 2021.

[56] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021.

[57] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017.

[58] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, pages 289–305, 2018.